

# MULTIPLE TESTING UNDER DEPENDENCE WITH APPROXIMATE CONDITIONAL LIKELIHOOD

Sairam Rayaprolu and Zhiyi Chi<sup>1</sup>

## Abstract

Statistical dependence in data poses a significant challenge to the stability of large scale multiple testing. This makes the variances of the number of true discoveries and the false discovery proportion crucial considerations in the false discovery rate control [15, 30, 31]. However, the statistical dependence structure of data is often unknown. We address the problem for the case of a stationary ergodic binary signal process embedded in noisy observations, where the distribution of the noise is known while that of the signal process is unknown. A novel aspect of our approach is the approximation of the conditional likelihoods of signals at individual sites given the data by incorporating nearby observations as well as estimates of low order moments obtained from a noiseless sample of the signal process. Simulations are carried out to show the stability as well as validity and power of the approach. The simulations rely on sampling of hidden Markov dependence structures as random matrices with specified stationary distribution and lower bounds on the top absolute eigenvalues, which is of interest in its own right.

*Keywords and phrases:* multiple hypothesis testing, FDR, HMM

*MSC 2010 subject classifications:* Primary 62H15, 62M02, 62M07

## 1 Introduction

Simultaneous testing of multiple hypotheses, or multiple testing for short, is of great importance to modern data science; see [16] for a broad overview. A great deal of research has been focused on controlling the false discovery rate (FDR) [2]. Since its introduction, the FDR has been widely used in applications where screening of massive data sets has to be carried out.

The outcomes of multiple testing are summarized below. In particular,  $V$  is known as the number of Type I errors or false discoveries,  $T$  the number of Type II errors or false nondiscoveries,  $S$  the number of true discoveries,  $R$  the total number of discoveries, and  $A$  the total number of nondiscoveries.

Hypothesis	Accept	Reject	Total
True null	$U$	$V$	$m_0$
False null	$T$	$S$	$m_1$
	$A$	$R$	$m$

By definition, the FDR is the expectation of  $V/(R \vee 1)$ , known as the false discovery proportion (FDP). Assuming independence of test statistics, the FDR can be controlled by the well-known Benjamini-Hochberg (BH) procedure [2, 41]. The BH procedure and its underlying philosophy have inspired much research on multiple testing. On the other hand, it can be argued that most data subject to multiple testing have certain statistical dependence. Under various assumptions of dependence, the BH procedure is still valid, i.e., able to control the FDR [3, 10, 35, 36, 43]. However, validity is mostly, though not entirely, concerned with Type I errors. To evaluate the performance of multiple testing, there are other important aspects that need to be taken into account as well. One is the control of Type II errors. A common measure for this is power, define as  $E(S)$  or  $E(S/m_1)$  depending on application. Another measure in similar spirit as the FDR is the

<sup>1</sup>Department of Statistics, University of Connecticut, 215 Glenbrook Road, U-4120, Storrs, CT 06269, sairam-ray@gmail.com and zhiyi.chi@uconn.edu.

false nondiscovery rate  $E[T/(A \vee 1)]$  [20]. Another important aspect of performance is stability, i.e., variation of the outcomes of multiple testing [22]. When several aspects are taken into account, in the presence of substantial correlation, the overall performance of the BH procedure is known to deteriorate, sometimes severely [14, 15, 30–32].

A natural idea to tackle statistical dependence is to explicitly model it for the underlying data. A few general approaches that have been developed to accomplish this rely upon dependence kernels and linear latent factor models [18, 19, 27]. Such approaches have had some success in improving power and stability in gene expression studies and brain imaging [18]. However, a latent dependence structure is not always readily identifiable and further may not always be accurately captured by a linear model. To balance the control of Type I and Type II errors, Sun and Cai [42] proposed to exploit the dependence structure of hypotheses from a decision-theoretical point of view and considered a hidden Markov model (HMM). Let  $H_1, \dots, H_m$  be a set of hypotheses and  $\eta_j = 0$  if  $H_j$  is true, and 1 otherwise. Assume that  $\boldsymbol{\eta} = (\eta_1, \dots, \eta_m)$  form a Markov chain and  $\mathbf{X} = (X_1, \dots, X_m)$  is a set of observations that are independent conditional on  $\boldsymbol{\eta}$ . Then they showed that a thresholding procedure for the conditional likelihoods  $P(\eta_j = 1 | \mathbf{X})$  minimizes  $E(\lambda V + T)$ . From a Bayesian perspective, the Markov chain is a prior on  $\boldsymbol{\eta}$  and  $P(\eta_j = 1 | \mathbf{X})$  the posterior likelihood for  $H_j$  being true. In an earlier work [37], it was shown that for any prior on  $\boldsymbol{\eta}$  and any type of data  $\mathbf{X}$ , the following BH-type procedure, which will be referred to as the Bayes BH procedure, controls the FDR at a fixed target level  $\alpha \in (0, 1)$ .

#### Bayes BH Procedure

1. Sort  $P(\eta_j = 1 | \mathbf{X})$  in ascending order as  $q_1 \leq q_2 \leq \dots \leq q_m$ .
2. If  $q_m < 1 - \alpha$ , then accept all  $H_j$ , otherwise reject all  $H_j$  with  $P(\eta_j = 1 | \mathbf{X}) \geq q_R$ , where  $R = \max\{1 \leq k \leq m : k^{-1} \sum_{j=1}^k q_{m-j+1} \geq 1 - \alpha\}$ .

It was shown in [9] that the Bayes BH procedure is power optimal *a posteriori*, in the sense that it has the largest  $E(S | \mathbf{X})$  among all procedures with  $E(\text{FDP} | \mathbf{X}) \leq \alpha$ . For comparison and later use, the BH procedure is based on  $p$ -values of hypotheses and can be described as follows [2, 38, 41].

#### BH Procedure

1. Sort the  $p$ -values in ascending order as  $p_1 \leq p_2 \leq \dots \leq p_m$ .
2. If  $p_1 > \alpha/m$ , then accept all  $H_j$ , otherwise reject all  $H_j$  with  $p_j \leq p_R$ , where  $R = \max\{1 \leq k \leq m : p_k \leq k\alpha/m\}$ .

The above results highlight the importance of  $P(\eta_j = 1 | \mathbf{X})$  in multiple testing. However, to precisely evaluate the conditional likelihoods either by closed-form calculation or simulation, it is necessary to elaborate on the joint dependence structure of  $\boldsymbol{\eta}$  and  $\mathbf{X}$ , which is typically impossible. In addition, from a signal processing point of view, multiple testing becomes useful typically when signals are relatively weak. We therefore wish to approximate  $P(\eta_j = 1 | \mathbf{X})$  using partial information on the dependence structure that is easy to infer, with reasonable accuracy especially when the signal is relatively weak.

In [9], under the HMM model,  $P(\eta_j | \mathbf{X})$  was approximated by second order Taylor series in signal strength. In this paper, the approach is extended to hypotheses with unknown dependence structure. Let  $(H_t)_{t \in T}$  be a set of hypotheses indexed by a finite set  $T$ , which can be as simple as  $\{1, \dots, m\}$ , or endowed with certain topology to incorporate, say spatial, relationships between the hypotheses. For each  $t \in T$ , let  $\eta_t = 0$  if  $H_t$  is true and 1 otherwise. In the most general setting,  $\mathbf{X}$  can be of any type. We assume that conditional on  $\boldsymbol{\eta}$ ,  $\mathbf{X}$  has a probability density of the form

$$\rho(\mathbf{x} | \boldsymbol{\eta}) = e^{q(\mathbf{x}, \boldsymbol{\eta})}. \quad (1)$$

The parameter  $\epsilon > 0$  in (1) will be referred to as “signal strength”. Observe that as  $\epsilon = 0$ ,  $\mathbf{X}$  is independent of  $\boldsymbol{\eta}$ , and hence contains no information about the hypotheses.

In Section 2, we approximate  $P(\eta_t = 1 \mid \mathbf{X})$  using second order Taylor series in  $\epsilon$ . The method, however, can attain higher order Taylor series as well. The approximation is valid when signal strength is low, precisely when the power of multiple testing suffers the most. When the effect of the hypotheses on the data is “localized”, i.e.,  $\mathbf{X} = (X_t)_{t \in T}$  with  $X_t$  being independent conditionally on  $\boldsymbol{\eta}$ , the approximation takes a simpler form. It only needs as inputs the first three moments of  $\boldsymbol{\eta}$  and the distribution of  $X_t$  conditional on  $\eta_t$  for individual  $t$ , as opposed to information on how the whole  $\mathbf{X}$  depends on  $\boldsymbol{\eta}$ . We will argue that in many settings of signal processing, both the moments of  $\boldsymbol{\eta}$  and the local dependence may be inferred with reasonable precision. Results of simulations and computations for the approximated conditional likelihoods are presented in Section 3.

*Notation.* We will treat each  $(x_t)_{t \in T}$  as a column vector. Thus, for  $\mathbf{u} = (u_t)_{t \in T}$  and  $\mathbf{v} = (v_t)_{t \in T}$ ,  $\mathbf{u}^\mathbf{t} \mathbf{v} = \sum_{t \in T} u_t v_t$ . If  $p$  is a function on  $\mathbb{R}$ , then  $p(\mathbf{x})$  is a shorthand for  $(p(x_t))_{t \in T}$ .  $(M_{st})_{s, t \in T}$  denotes a matrix with rows indexed by  $s$  and columns indexed by  $t$ . Finally,  $\text{diag}(\mathbf{x})$  denotes the diagonal matrix  $(M_{st})_{s, t \in T}$  with  $M_{tt} = x_t$  and  $M_{st} = 0$  for  $s \neq t$ .

## 2 Approximation of conditional likelihoods

### 2.1 A general Taylor series approximation

While our objective is to approximate  $P(\eta_t = 1 \mid \mathbf{X})$ , the conditional log-likelihood ratios

$$r_t(\mathbf{X}) = \ln \frac{P(\eta_t = 1 \mid \mathbf{X})}{P(\eta_t = 0 \mid \mathbf{X})}, \quad t \in T$$

are more convenient to work with. Each  $r_t(\mathbf{X})$  is simply the logit transform of  $P(\eta_t = 1 \mid \mathbf{X})$ . Conversely, the latter is the logistic transform of the former, i.e.,

$$P(\eta_t = 1 \mid \mathbf{X}) = \frac{1}{1 + \exp\{-r_t(\mathbf{X})\}}.$$

Henceforth, quantities derived from taking expectation conditional on  $\eta_t = i$  will be indexed by  $i$  and  $t$ , e.g.,

$$\mathbb{E}_{it}(\cdot) = \mathbb{E}(\cdot \mid \eta_t = i), \quad \rho_{it}(\cdot) = \rho(\cdot \mid \eta_t = i), \quad \text{Cov}_{it}(\cdot) = \text{Cov}(\cdot \mid \eta_t = i).$$

By Bayes rule, for  $t \in T$  and  $i = 0, 1$ ,  $P(\eta_t = i \mid \mathbf{X}) \propto P(\eta_t = i) \rho_{it}(\mathbf{X})$ . Then

$$r_t(\mathbf{X}) = \ln \frac{P(\eta_t = 1)}{P(\eta_t = 0)} + \ln \frac{\rho_{1t}(\mathbf{X})}{\rho_{0t}(\mathbf{X})} \quad (2)$$

and it is essential to evaluate the second term on the r.h.s. of (2). By conditioning,

$$\rho_{it}(\mathbf{X}) = \sum_{\boldsymbol{\sigma}: \sigma_t = i} P_{it}(\boldsymbol{\eta} = \boldsymbol{\sigma}) \rho(\mathbf{X} \mid \boldsymbol{\eta} = \boldsymbol{\sigma}). \quad (3)$$

**Lemma 2.1.** *Let  $\boldsymbol{\eta}$  be a binary process on  $T$ . Given  $g \in C^2(\mathbb{R}^T, \mathbb{R})$ , denote its gradient and Hessian by  $g'$  and  $g''$ , respectively. Then as  $\epsilon \rightarrow 0$ ,*

$$\ln \mathbb{E}[e^{g(\epsilon \boldsymbol{\eta})}] = g(\mathbf{0}) + [g'(\mathbf{0})^\mathbf{t} \mathbb{E}(\boldsymbol{\eta})] \epsilon + \frac{1}{2} [\mathbb{E}(\boldsymbol{\eta}^\mathbf{t} g''(\mathbf{0}) \boldsymbol{\eta}) + g'(\mathbf{0})^\mathbf{t} \text{Cov}(\boldsymbol{\eta}) g'(\mathbf{0})] \epsilon^2 + o(\epsilon^2). \quad (4)$$

Moreover, if  $\boldsymbol{\pi} = (\pi_t)_{t \in T}$  and  $\mathbf{J} = (J_{st})_{s, t \in T}$ , with  $\pi_t = P(\eta_t = 1)$  and  $J_{st} = P(\eta_s = \eta_t = 1)$ , then

$$\mathbb{E}(\boldsymbol{\eta}) = \boldsymbol{\pi}, \quad \mathbb{E}(\boldsymbol{\eta}^\mathbf{t} g''(\mathbf{0}) \boldsymbol{\eta}) = \text{Tr}(g''(\mathbf{0}) \mathbf{J}), \quad \text{Cov}(\boldsymbol{\eta}) = \mathbf{J} - \boldsymbol{\pi} \boldsymbol{\pi}^\mathbf{t}. \quad (5)$$

*Proof.* Eq. (4) is well known and can be checked by standard calculation. Since  $\eta_t$  is binary,  $E(\eta_t) = P(\eta_t = 1) = \pi_t$ , yielding the first equation in (5). Likewise,  $E(\boldsymbol{\eta}\boldsymbol{\eta}^\dagger) = \mathbf{J}$ . Then the second equation in (5) follows from

$$E(\boldsymbol{\eta}^\dagger g''(\mathbf{0})\boldsymbol{\eta}) = E(\text{Tr}(\boldsymbol{\eta}^\dagger g''(\mathbf{0})\boldsymbol{\eta})) = E(\text{Tr}(g''(\mathbf{0})\boldsymbol{\eta}\boldsymbol{\eta}^\dagger)) = \text{Tr}(E(g''(\mathbf{0})\boldsymbol{\eta}\boldsymbol{\eta}^\dagger)) = \text{Tr}(g''(\mathbf{0})\mathbf{J}).$$

The last equation in (5) is also easy to verify.  $\square$

**Theorem 2.2.** *Let  $q$  be as in (1). Suppose that for each  $\mathbf{x}$ ,  $q(\mathbf{x}, \boldsymbol{\theta})$  as a function of  $\boldsymbol{\theta} \in \mathbb{R}^T$  belongs to  $C^2$ . Let  $\gamma(\mathbf{x})$  and  $H(\mathbf{x})$  be the gradient and Hessian of  $q(\mathbf{x}, \boldsymbol{\theta})$  at  $\boldsymbol{\theta} = \mathbf{0}$ , respectively. Then*

$$\begin{aligned} r_t(\mathbf{X}) = & \ln \frac{P(\eta_t = 1)}{P(\eta_t = 0)} + \gamma(\mathbf{X})^\dagger [E_{1t}(\boldsymbol{\eta}) - E_{0t}(\boldsymbol{\eta})]\epsilon \\ & + \frac{1}{2} [E_{1t}(\boldsymbol{\eta}^\dagger H(\mathbf{X})\boldsymbol{\eta}) - E_{0t}(\boldsymbol{\eta}^\dagger H(\mathbf{X})\boldsymbol{\eta})]\epsilon^2 \\ & + \frac{1}{2} \gamma(\mathbf{X})^\dagger [\text{Cov}_{1t}(\boldsymbol{\eta}) - \text{Cov}_{0t}(\boldsymbol{\eta})] \gamma(\mathbf{X}) \epsilon^2 + o_{t,\mathbf{X}}(\epsilon^2) \end{aligned} \quad (6)$$

as  $\epsilon \rightarrow 0$ , where  $o_{t,\mathbf{X}}(\cdot)$  means that the implicit constant depends on  $t$  and  $\mathbf{X}$ .

*Proof.* From (3) and the assumption,  $\rho_{it}(\mathbf{X}) = E_{it}[\rho(\mathbf{X} | \boldsymbol{\eta})] = E_{it}[e^{q(\mathbf{X}, \epsilon \boldsymbol{\eta})}]$ . Note that  $\mathbf{X}$  is fixed. Applying Lemma 2.1 to  $g(\cdot) := q(\mathbf{X}, \cdot)$ , for  $i = 0, 1$ ,

$$\begin{aligned} \ln \rho_{it}(\mathbf{X}) = & q(\mathbf{X}, \mathbf{0}) + [\gamma(\mathbf{X})^\dagger E_{it}(\boldsymbol{\eta})]\epsilon \\ & + \frac{1}{2} [E_{it}(\boldsymbol{\eta}^\dagger H(\mathbf{X})\boldsymbol{\eta}) + \gamma(\mathbf{X})^\dagger \text{Cov}_{it}(\boldsymbol{\eta}) \gamma(\mathbf{X})]\epsilon^2 + o_{t,\mathbf{X}}(\epsilon^2). \end{aligned}$$

This together with (2) yields the claimed expansion.  $\square$

## 2.2 Localized dependence of data on hypotheses

We now consider the case where the dependence of  $\mathbf{X}$  on  $\boldsymbol{\eta}$  is localized. Specifically, suppose  $\mathbf{X} = (X_t)_{t \in T}$  and the function  $q$  in (1) can be written as

$$q(\mathbf{x}, \boldsymbol{\theta}) = \sum_t q_t(x_t, \theta_t). \quad (7)$$

It follows that  $X_t$  are independent conditionally on  $\boldsymbol{\eta}$ . We have the following result.

**Corollary 2.3.** *Suppose that for each  $x$  and  $t \in T$ ,  $q_t(x, \theta)$  as a function of  $\theta \in \mathbb{R}$  belongs to  $C^2$ . Let  $\gamma_t(x)$  and  $k_t(x)$  be the first and second derivatives of  $q_t(x, \theta)$  at  $\theta = 0$ , respectively. Denote  $\gamma(\mathbf{x}) = (\gamma_t(x_t))_{t \in T}$  and  $k(\mathbf{x}) = (k_t(x_t))_{t \in T}$ . Then as  $\epsilon \rightarrow 0$ ,*

$$\begin{aligned} r_t(\mathbf{X}) = & \ln \frac{P(\eta_t = 1)}{P(\eta_t = 0)} + \gamma(\mathbf{X})^\dagger [E_{1t}(\boldsymbol{\eta}) - E_{0t}(\boldsymbol{\eta})]\epsilon \\ & + \frac{1}{2} k(\mathbf{X})^\dagger [E_{1t}(\boldsymbol{\eta}) - E_{0t}(\boldsymbol{\eta})]\epsilon^2 \\ & + \frac{1}{2} \gamma(\mathbf{X})^\dagger [\text{Cov}_{1t}(\boldsymbol{\eta}) - \text{Cov}_{0t}(\boldsymbol{\eta})] \gamma(\mathbf{X}) \epsilon^2 + o_{t,\mathbf{X}}(\epsilon^2). \end{aligned} \quad (8)$$

*Proof.* Comparing (6) and (8), it suffices to show

$$E_{1t}(\boldsymbol{\eta}^\dagger H(\mathbf{X})\boldsymbol{\eta}) - E_{0t}(\boldsymbol{\eta}^\dagger H(\mathbf{X})\boldsymbol{\eta}) = k(\mathbf{X})^\dagger [E_{1t}(\boldsymbol{\eta}) - E_{0t}(\boldsymbol{\eta})].$$

From (7), it is seen that  $H(\mathbf{x}) = \text{diag}(k(\mathbf{x}))$ . Then for  $i = 0, 1$ ,

$$\begin{aligned} E_{it}(\boldsymbol{\eta}^t H(\mathbf{X}) \boldsymbol{\eta}) &= E_{it}(\text{Tr}(\boldsymbol{\eta}^t H(\mathbf{X}) \boldsymbol{\eta})) = E_{it}(\text{Tr}(\text{diag}(k(\mathbf{X})) \boldsymbol{\eta} \boldsymbol{\eta}^t)) \\ &= \sum_s k_s(x_s) P_{it}(\eta_s = 1) = k(\mathbf{x})^t E_{it}(\boldsymbol{\eta}), \end{aligned}$$

and hence the desired identity.  $\square$

In the expansion (8), since  $E_{it}(\boldsymbol{\eta})$  is the vector consisting of  $P(\eta_s = 1 | \eta_t = i)$ ,  $s \in T$ , it is determined by the first and second moments of  $\boldsymbol{\eta}$ . Likewise,  $\text{Cov}_{it}(\boldsymbol{\eta})$  is determined by the first three moments of  $\boldsymbol{\eta}$ . On the other hand,  $\gamma_t(x)$  and  $k_t(x)$  are determined by how  $X_t$  depends on  $\eta_t$ , regardless of the statistical dependence at other  $s \in T$ . Under many settings of signal processing, these quantities can be estimated directly. One can regard  $\boldsymbol{\eta}$  as a message sent through a noisy environment, and  $\mathbf{X}$  the noise-corrupted message being received. Before the communication formally starts, one can first estimate the moments of the distribution of  $\boldsymbol{\eta}$ . For example, if  $\boldsymbol{\eta}$  is a long contiguous segment of a stationary and ergodic binary process, then the moments may be estimated using a single observation of  $\boldsymbol{\eta}$  by the law of large numbers. If in addition, there is no significant long range dependence in the distribution of  $\boldsymbol{\eta}$ , then the evaluation of the moments, say  $E(\eta_0 \eta_t)$ , can be restricted to  $t$  within an appropriately chosen radius from 0, which significantly reduces the number of moments to be estimated. On the other hand, one can use pre-selected test signals to estimate the dependence of  $X_t$  on  $\eta_t$ . If for all  $t$ , the dependence is the same, i.e., the function  $q_t$  is the same for all  $t$ , then an estimation may be obtained from a single observation of the pair  $(\boldsymbol{\eta}, \mathbf{X})$ , again by the law of large numbers.

Sometimes it is more convenient to express  $X_t$  as a result of interaction between  $\eta_t$  and  $Z_t$ , where  $Z_t$  are i.i.d. and regarded as noise. Specifically, suppose

$$\begin{aligned} X_t &= \varphi_t(\epsilon \eta_t, Z_t), \quad \text{with } \mathbf{Z} = (Z_t)_{t \in T} \text{ independent of } \boldsymbol{\eta} \\ &\text{and } Z_t \text{ i.i.d. with density } e^{h(z)}. \end{aligned} \tag{9}$$

Then the distribution of  $X_t$  conditional on  $\eta_t$  can be derived from the distribution of  $Z_t$  and the functional form of  $\varphi_t$ . We next consider two important examples.

**Example 2.4** (Additive noise). Let  $\varphi_t(u, z) = u + z$  for all  $t$ . We then find that

$$\rho(\mathbf{X} | \boldsymbol{\eta}) = \prod_t \exp(h(X_t - \epsilon \eta_t)) = e^{q(\mathbf{X}, \epsilon \boldsymbol{\eta})},$$

with  $q(\mathbf{x}, \mathbf{u}) = \sum_t h(x_t - u_t)$ . Then  $\gamma(\mathbf{x}) = -h'(\mathbf{x})$  and  $H(\mathbf{x}) = \text{diag}(h''(\mathbf{x}))$ , so by Theorem 2.2,

$$\begin{aligned} r_t(\mathbf{X}) &= \ln \frac{P(\eta_t = 1)}{P(\eta_t = 0)} - h'(\mathbf{X})[E_{1t}(\boldsymbol{\eta}) - E_{0t}(\boldsymbol{\eta})]\epsilon \\ &\quad + \frac{1}{2} h'(\mathbf{X})^t [\text{Cov}_{1t}(\boldsymbol{\eta}) - \text{Cov}_{0t}(\boldsymbol{\eta})] h'(\mathbf{X}) \epsilon^2 \\ &\quad + \frac{1}{2} h''(\mathbf{X})^t [E_{1t}(\boldsymbol{\eta}) - E_{0t}(\boldsymbol{\eta})] \epsilon^2 + o_{t, \mathbf{X}}(\epsilon^2), \quad \text{as } \epsilon \rightarrow 0. \end{aligned} \tag{10}$$

**Example 2.5** (Multiplicative noise). Let  $\varphi_t(u, z) = ze^{-u}$  for all  $t$ . Then,

$$\rho(\mathbf{X} | \boldsymbol{\eta}) = \prod_t \exp(\epsilon \eta_t) \exp(h(X_t \exp(\epsilon \eta_t))) = e^{q(\mathbf{X}, \epsilon \boldsymbol{\eta})},$$

with  $q(\mathbf{x}, \mathbf{u}) = \sum_t [u_t + h(x_t e^{u_t})]$ . Define  $g(x) = 1 + xh'(x)$  and  $u(x) = xh'(x) + x^2h''(x)$ . It is easy to get  $\gamma(\mathbf{x}) = g(\mathbf{x})$  and  $H(\mathbf{x}) = \text{diag}(u(\mathbf{x}))$ . Then, as in the additive case,

$$\begin{aligned} r_t(\mathbf{X}) &= \ln \frac{P(\eta_t = 1)}{P(\eta_t = 0)} + g(\mathbf{X})[E_{1t}(\boldsymbol{\eta}) - E_{0t}(\boldsymbol{\eta})]\epsilon \\ &\quad + \frac{1}{2}g(\mathbf{X})^\dagger [\text{Cov}_{1t}(\boldsymbol{\eta}) - \text{Cov}_{0t}(\boldsymbol{\eta})]g(\mathbf{X})\epsilon^2 \\ &\quad + \frac{1}{2}u(\mathbf{X})^\dagger [E_{1t}(\boldsymbol{\eta}) - E_{0t}(\boldsymbol{\eta})]\epsilon^2 + o_{t,\mathbf{X}}(\epsilon^2), \quad \text{as } \epsilon \rightarrow 0. \end{aligned}$$

### 3 Numerical experiments

In this section, we report simulation studies to assess multiple testing based on approximated conditional likelihoods in the presence of dependence. We compare the Bayes BH procedure and the BH procedure. Recall that the former is based on the approximated conditional likelihoods of hypotheses given the data while the latter is based on marginal  $p$ -values of hypotheses.

#### 3.1 Sampling of chain of null hypotheses

In the simulations,  $\boldsymbol{\eta} = (\eta_1, \dots, \eta_m)$  is a long binary hidden Markov chain. This setting is very flexible since hidden Markov models are dense among essentially all finite-state stationary processes [26]. The chain is sampled as follows.

1. Specify a finite state space  $E$ , a nonempty strict subset  $F$  of  $E$ , and a function  $\tau : E \rightarrow \{0, 1\}$  that maps  $F$  to 0 and  $E \setminus F$  to 1.
2. Randomly sample a probability vector  $\boldsymbol{\pi} = (\pi_s)_{s \in E}$  and a transition matrix  $\mathbf{P}$  on the states of  $E$  such that

$$\boldsymbol{\pi}^\dagger \mathbf{P} = \boldsymbol{\pi}^\dagger \tag{11}$$

3. Simulate a stationary Markov chain  $\mathbf{M} = (M_1, \dots, M_m)$  on  $E$  with stationary distribution  $\boldsymbol{\pi}$  and transition matrix  $\mathbf{P}$ . Set  $\eta_t = \tau(M_t)$  for each  $t$ . We will abbreviate the transform as

$$\boldsymbol{\eta} = \tau(\mathbf{M})$$

and refer to  $\mathbf{M}$  as the “parent” Markov chain.

In all the simulations of the study,  $E = \{1, 2, 3, 4, 5\}$ ,  $F = \{1, 2\}$ , and  $m = 10^5$ . The function  $\tau$  is used only for the simulation of  $\boldsymbol{\eta}$  and not available for multiple testing. Thus, each realization of  $\boldsymbol{\eta}$  is regarded as a large set of hypotheses with an unknown dependence structure. To avoid artifacts that may confound the comparison of multiple testing procedures, both  $\boldsymbol{\pi}$  and  $\mathbf{P}$  are randomly sampled. Denoting by  $\psi$  the proportion of false nulls, we need  $\boldsymbol{\pi}$  to be such that the proportion of 1’s in  $\boldsymbol{\eta}$  is  $\psi$ . This is achieved by setting  $\pi_s = (1 - \psi)\zeta_s / \sum_{s \in F} \zeta_s$  for  $s \in F$  and  $\pi_s = \psi\zeta_s / \sum_{s \notin F} \zeta_s$  for  $s \notin F$ , with  $\zeta_s$  i.i.d.  $\sim$  uniformly distributed on  $(0, 1)$ .

Given  $\boldsymbol{\pi}$ , the key component of the simulation is the transition matrix  $\mathbf{P}$ , which is to be sampled from positive matrices, i.e., matrices with all entries being positive. Consequently,  $\mathbf{M}$  is irreducible and aperiodic. In addition to the constraint (11), since we need to control the strength of dependence within  $\boldsymbol{\eta}$ , it is necessary that the mixing rate of  $\mathbf{M}$  is controllable. As is well known, 1 is an eigenvalue of  $\mathbf{P}$  with multiplicity one, and all the other eigenvalues of  $\mathbf{P}$  have absolute values, i.e., eigenmoduli, strictly less than 1. Sort the eigenmoduli in decreasing order as  $1 = \varrho_1(\mathbf{P}) > \varrho_2(\mathbf{P}) \geq \dots \geq \varrho_d(\mathbf{P}) \geq 0$ . It is well known that the second largest eigenmodulus is an

important parameter in determining the mixing rate of  $\mathbf{M}$  [33], with small value of  $\varrho_2(\mathbf{P})$  giving rise to high mixing rate and hence weak dependence within the chain. However, the mixing rate is not determined by  $\varrho_2(\mathbf{P})$  alone and actually is affected by the spectrum of the transition matrix in a surprisingly complicated way [11, 28]. The difficulty to control the strength of dependence within  $\boldsymbol{\eta}$  is further exacerbated by the fact that it is a many-to-one transform of  $\mathbf{M}$ , so itself is not Markov. Thus in addition to second largest eigenmodulus, we also consider the third largest one. Since the transition matrices simulated in the numerical experiments are all  $5 \times 5$ , these two parameters provide significant information about their spectra. We next describe a heuristic procedure to sample  $\mathbf{P}$  that has prescribed properties.

### 3.2 Sampling of random transition matrices

The ensemble of  $d \times d$  uniformly distributed transition matrices, known as the Dirichlet Markov Ensemble, can be represented by the uniform distribution on  $\Delta_d^d$ , the  $d$ -fold Cartesian product of  $\Delta_d = \{(x_1, \dots, x_d) \in [0, 1]^d : x_1 + \dots + x_d = 1\}$ . The support of this ensemble is a convex  $d(d-1)$  dimensional polytope. Uniform sampling of  $\mathbf{P}$  from the ensemble is easy because the rows of  $\mathbf{P}$  are i.i.d.  $\sim (\xi_1, \dots, \xi_d)/(\xi_1 + \dots + \xi_d)$  with  $\xi_i$  i.i.d. exponentially distributed; see [7] for more detail. However, in our study, we need to sample from the  $(d-1)^2$  dimensional polytope of transition matrices that have a prescribed stationary distribution  $\boldsymbol{\pi}$ . To the best of our knowledge, exact uniform sampling from this polytope is unknown for general  $\boldsymbol{\pi}$  [6]. An alternative to exact sampling is to use asymptotically exact methods. For example, [12] constructs a Markov chain whose stationary distribution is the uniform distribution on a convex polytope. In [8], a Gibbs sampler is provided for the uniform distribution on the set of doubly stochastic matrices, i.e., transition matrices whose transposes are also transition matrices. However, doubly stochastic matrices are restrictive for our study, as their stationary distribution is always the uniform one on  $1, \dots, d$ . To sample from the polytope of transition matrices with a specified stationary distribution, we use a heuristic described below. It should be pointed out that matrices sampled this way are not uniformly distributed on the polytope associated with  $\boldsymbol{\pi}$ . Also, the complexity of the sampling increases fast as  $d$  increases, as the dimension of the polytope grows at the order of  $d^2$ . We mention that the properties of a “typical” matrix from the uniform distribution on a closely related polytope of matrices was studied in [1].

Sinkhorn’s algorithm [39, 40] is an iterative scaling procedure that maps each positive matrix, not necessarily square, to a unique positive matrix with prescribed row and column sums. Its scheme is simply to scale the rows of a matrix to obtain the prescribed row sums, then scale the columns of the resulting matrix to obtain the prescribed column sums, and so on until convergence. The algorithm has been extended in many works (cf. [5, 13, 23, 25, 29, 34]), and its geometric rate of convergence and validity for irreducible nonnegative matrices have been established [17]. We will use the following version of the algorithm [40], which, given a  $d$ -dimensional probability vector  $\boldsymbol{\pi}$  with all entries being positive, samples a positive matrix  $\mathbf{A}$  whose row sum and column sum vectors are both equal to  $\boldsymbol{\pi}$  up to a prescribed precision level. Consequently, the final output  $\mathbf{P}$  is a transition matrix with stationary distribution (approximately) equal to  $\boldsymbol{\pi}$  [24].

#### Sinkhorn’s Algorithm

Input:  $\boldsymbol{\pi} \in \text{Interior}(\Delta_d)$

Sample  $\mathbf{A}$  uniformly from the set of  $d \times d$  transition matrices. ( $\mathbf{A}$  is positive w.p. 1.)

While  $\|\mathbf{A}\mathbf{1}_d - \boldsymbol{\pi}\| > \epsilon$  or  $\|\mathbf{1}_d^t \mathbf{A} - \boldsymbol{\pi}^t\| > \epsilon$

$$\mathbf{D}^L \leftarrow [\text{diag}(\mathbf{A}\mathbf{1}_d)]^{-1} \text{diag}(\boldsymbol{\pi})$$

$$\mathbf{A} \leftarrow \mathbf{D}^L \mathbf{A}$$

$$\mathbf{D}^R \leftarrow [\text{diag}(\mathbf{1}_d^t \mathbf{A})]^{-1} \text{diag}(\boldsymbol{\pi})$$



$\mathbf{A} \leftarrow \mathbf{A}\mathbf{D}^R$   
 Return  $\mathbf{P} = \text{diag}(\boldsymbol{\pi})^{-1}\mathbf{A}$

Sinkhorn’s algorithm partitions the set of irreducible nonnegative matrices into equivalent classes, each consisting of matrices that converge to the same transition matrix with stationary distribution  $\boldsymbol{\pi}$ . However, since there is no closed form characterization of the equivalence, the sampling distribution of the algorithm in general is unknown. For the special case of  $\boldsymbol{\pi} = \mathbf{1}_d/d$ , the sampling distribution is in fact not uniform but “locally flat” at the center of the polytope of doubly stochastic matrices [6].

Despite its intractable sampling distribution, Sinkhorn’s algorithm has several advantages. First, it allows prescribing the stationary distribution of  $\mathbf{P}$  exactly. Second, it is computationally tractable due to its scalability to high dimensions, geometric convergence, and simplicity. Third, it allows generation of irreducible transition matrices containing one or more 0’s by starting with a nonnegative matrix with 0’s in exactly the same entries [5]. Consequently, it is applicable to simulation of Markov chains on strongly connected finite directed graphs. It is known that under the uniform distribution on  $\Delta_d^d$ , all the eigenvalues of a  $d \times d$  transition matrix, except 1, converge in probability to 0 as  $d \rightarrow \infty$  [4, 21]. Our numerical experiments indicate that this is also the case under the sampling distribution of Sinkhorn’s algorithm.

To specify the size of the leading eigenmoduli of  $\mathbf{P}$ , choose  $\lambda \in [0, 1)$  as a lower bound on  $\varrho_2(\mathbf{P})$  and  $0 \leq \mu < \lambda$  as a lower bound on  $\varrho_3(\mathbf{P})$ . For  $\lambda = 0$ , we simply set  $\mathbf{P} = \mathbf{1}\boldsymbol{\pi}^t$ , while for  $\lambda > 0$  and  $\mu \geq 0$  we repeatedly sample  $\mathbf{P}$  by Sinkhorn’s algorithm until  $\varrho_2(\mathbf{P}) \geq \lambda$  and  $\varrho_3(\mathbf{P}) \geq \mu$ .

### 3.3 Comparing Bayes BH and BH procedures

After  $\boldsymbol{\pi}$  and  $\mathbf{P}$  are sampled and fixed, the simulation proceeds in two steps. First, one realization of  $\boldsymbol{\eta}$  is sampled and used to estimate  $E(\eta_0)$ ,  $E(\eta_i | \eta_0)$ , and  $E(\eta_i \eta_j | \eta_0)$ . We denote this realization by  $\boldsymbol{\theta}$ . Then, an independent realization of  $\boldsymbol{\eta}$ , still denoted by  $\boldsymbol{\eta}$ , is sampled and used to generate data  $\mathbf{X} = \epsilon \boldsymbol{\eta} + \mathbf{Z}$ , where the entries of  $\mathbf{Z}$  are i.i.d.  $\sim N(0, 1)$  and  $\epsilon > 0$  is the signal strength. In the simulation,  $\boldsymbol{\eta}$  can be viewed as a signal that needs to be sent through a noisy environment. To do this, we first estimate the moments and conditional moments of the signal up to order two using a sample  $\boldsymbol{\theta}$ . Meanwhile, we estimate the distribution of the noise  $\mathbf{Z}$ , for example, by sending a long sequence of 0’s. In principle, we can have very good knowledge about the noise environment by repeatedly testing it. Then, based on the estimates, multiple testing is used to recover new signals sent through the environment.

Given  $\mathbf{X}$ ,  $P(\eta_t = 1 | \mathbf{X})$  are approximately evaluated using the result in Example 2.4. To reduce computation, only  $E(\eta_s | \eta_t)$  and  $E(\eta_s \eta_u | \eta_t)$  with  $|s - t| \leq w$  and  $|u - t| \leq w$  are evaluated, where the “half window length”  $w \geq 0$  determines the number of moments that need to be incorporated in the approximation. All the other moments are treated as 0. In the absence of long range dependence,  $w$  can be small. The approximated  $P(\eta_t = 1 | \mathbf{X})$  are then used by the Bayes BH procedure. On the other hand, the marginal one-sided  $p$ -values based on individual  $X_t$  are used by the BH procedure.

Let  $\alpha$  be the nominal FDR control level for both the Bayes BH and the BH procedures. The Bayes BH procedure can estimate the population proportion of false nulls using the observed sample  $\boldsymbol{\theta}$ . However, the BH procedure does not rely on  $\boldsymbol{\theta}$ . To level the ground of comparison, the nominal FDR control level for the BH procedure is increased to

$$\tilde{\alpha} = \alpha / (1 - \tilde{\psi}) \quad \text{with } \tilde{\psi} = (\eta_1 + \dots + \eta_m) / m,$$

i.e.,  $\tilde{\psi}$  is the actual proportion of false nulls underlying the data  $\mathbf{X} = \epsilon \boldsymbol{\eta} + \mathbf{Z}$ . It is well known that the BH procedure under the augmented FDR control level is more powerful while still controlling



the FDR at the nominal level [2]. In contrast, the Bayes BH procedure has no access to statistics of  $\boldsymbol{\eta}$  or  $\mathbf{Z}$ . It uses  $\hat{\psi} = m^{-1} \sum \theta_t$  as the estimate of the proportion of false nulls, and derives all the estimates of moments and conditional moments from the observed sample  $\boldsymbol{\theta}$ .

For a single instance of multiple testing, its performance can be measured by the FDP and the number of true discoveries (NTD). Both measures are random variables as they are functions of the data as well as the procedure being used. Their population means and variances can be used to assess the overall performance of a multiple testing procedure, in particular, its validity, power, and stability. To estimate these parameters, the Bayes BH and the BH procedures are applied to multiple replications of data, all of which are generated with the same  $\boldsymbol{\eta}$  but with independent realizations of noise. To be specific, given  $\boldsymbol{\eta}$ , the following steps are taken.

1. Sample  $\mathbf{X} = \epsilon \boldsymbol{\eta} + \mathbf{Z}$ . At nominal FDR control level  $\alpha$ , apply the Bayes BH procedure to the approximated  $P(\eta_t = 1 | \mathbf{X})$ . On the other hand, at the augmented nominal FDR control level  $\tilde{\alpha}$ , apply the BH procedure to the marginal  $p$ -values of  $X_t$  under the null  $\eta_t = 0$ .
2. Repeat the above step  $n = 1000$  times. Compare the FDP and NTD of the Bayes BH procedure and those of the BH procedure, respectively, in terms of sample mean, sample standard error (SE), and empirical density.

### 3.4 Results

First, the main parameters used in the simulations are as follows; see Sections 3.1–3.3 for detail. The proportion of false nulls  $\psi = P(\eta_t = 1)$  takes values in  $\{0.05, 0.1\}$ , the signal strength  $\epsilon$  in  $\{0.75, 1, 1.25\}$ , and the nominal FDR control level  $\alpha$  in  $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ . For the purpose of these numerical experiments, an eigenmodulus is considered to be moderate if it about 0.5 to 0.6, large if above 0.9, and small if below 0.2. Given half window size  $w$ , only  $E(\eta_s | \eta_t)$ ,  $E(\eta_s \eta_u | \eta_t)$  with  $|s - t| \leq w$  and  $|u - t| \leq w$  are used to approximate  $P(\eta_t = 1 | \mathbf{X})$ , where  $\boldsymbol{\eta} = \tau(\mathbf{M})$  and  $\mathbf{X} = \epsilon \boldsymbol{\eta} + \mathbf{Z}$ . If  $\varrho_2(\mathbf{P}) > 0$ , then  $w = 3$ . On the other hand, if  $\varrho_2(\mathbf{P}) = 0$ , i.e.,  $\eta_s$  are i.i.d.  $\sim \pi$ , then  $w = 0$ . The length of  $\boldsymbol{\eta}$  is  $m = 10^5$ . Sample mean, sample SE, and empirical density of the outcomes of multiple testing are based on  $n = 1000$  replications of  $\mathbf{X}$  with  $\boldsymbol{\eta}$  being fixed.

#### 3.4.1 Large $\varrho_2$ and small $\varrho_3$

Table 1 displays the results of a set of simulations in which  $\varrho_2(\mathbf{P})$  and  $\varrho_3(\mathbf{P})$  are large and small respectively. The parent Markov chain  $\mathbf{M}$  has  $\varrho_2(\mathbf{P}) = 0.9074$ ,  $\varrho_3(\mathbf{P}) = 0.0645$ ,

$$\mathbf{P} = \begin{pmatrix} 0.9286 & 0.0122 & 0.0276 & 0.0063 & 0.0253 \\ 0.0150 & 0.8548 & 0.0651 & 0.0310 & 0.0340 \\ 0.0250 & 0.9629 & 0.0097 & 0.0008 & 0.0016 \\ 0.1653 & 0.7046 & 0.0955 & 0.0267 & 0.0079 \\ 0.2716 & 0.6988 & 0.0257 & 0.0007 & 0.0033 \end{pmatrix}, \quad (12)$$

$\pi = (0.3006, 0.5994, 0.0505, 0.0211, 0.0283)^t$ , and  $\psi = 0.1$ . Both the Bayes BH and the BH procedures control the FDR around the nominal level  $\alpha$ , with the former having somewhat smaller variance of the FDP. On the other hand, while both procedures have similar average NTD, the Bayes BH procedure has substantially less variation in the NTD. Thus, the approximate conditional probabilities provide better stability than the marginal  $p$ -values. In spite of the large  $\varrho_2(\mathbf{P})$ , the half window length  $w = 3$  seems to work well. Table 2 displays the results of another set of simulations in which  $\varrho_2(\mathbf{P})$  is even larger and  $\varrho_3(\mathbf{P})$  is still small. It shows similar phenomenon

Table 1: Comparison of the Bayes BH (BBH) and the BH procedures by sample means and variations of FDP and NTD in the presence large  $\varrho_2$  and small  $\varrho_3$ . The transition matrix of the parent Markov chain is (12) with  $\varrho_2 = 0.9074$ ,  $\varrho_3 = 0.0645$ , and  $\psi = 10\%$ .

		BBH	BH	BBH	BH
$\epsilon$	$\alpha$	$\hat{\mu}_{\text{FDP}}$		$\text{SE}(\hat{\mu}_{\text{FDP}})$	
0.75	0.4	0.41246	0.40276	0.069156	0.12349
0.75	0.5	0.50544	0.49928	0.030706	0.051165
1	0.1	0.0469	0.0999	0.1682	0.1633
1	0.15	0.1478	0.1465	0.1321	0.1324
1	0.2	0.2095	0.2043	0.0672	0.0995
1	0.2	0.2034	0.2010	0.0719	0.0968
1	0.3	0.3085	0.3027	0.0325	0.0422
1.5	0.1	0.10743	0.0994	0.016124	0.017342
1.5	0.2	0.20871	0.19912	0.010182	0.012625

$\epsilon$	$\alpha$	$\hat{\mu}_{\text{NTD}}$		$\text{SE}(\hat{\mu}_{\text{NTD}})$	
0.75	0.4	29.259	34.002	6.6888	23.973
0.75	0.5	134.2	131.19	13.801	57.623
1	0.1	1.0530	4.9680	1.4192	5.1406
1	0.15	8.0090	12.5270	3.8180	10.5292
1	0.2	28.7990	30.8470	6.6985	18.7630
1	0.2	24.7470	30.1690	6.1466	18.8499
1	0.3	137.0970	132.3660	13.4961	43.9223
1.5	0.1	358.46	322.19	22.156	46.92
1.5	0.2	1096.5	1028.1	37.538	73.39

observed in Table 1. In the simulations,  $\varrho_2(\mathbf{P}) = 0.9566$ ,  $\varrho_3(\mathbf{P}) = 0.1155$ ,

$$\mathbf{P} = \begin{pmatrix} 0.8993 & 0.0017522 & 0.0098301 & 0.031722 & 0.057401 \\ 0.015641 & 0.96471 & 0.0086498 & 0.0077613 & 0.0032366 \\ 0.66064 & 0.11572 & 0.075542 & 0.067761 & 0.080337 \\ 0.71019 & 0.090088 & 0.06356 & 0.13104 & 0.0051248 \\ 0.57536 & 0.025292 & 0.27334 & 0.10032 & 0.025692 \end{pmatrix} \quad (13)$$

with  $\boldsymbol{\pi} = (0.67187, 0.22813, 0.02414, 0.033354, 0.042506)^\top$ , and  $\psi = 0.1$ .

In addition to summary statistics such as sample mean and sample SE, density curves also show better stability of the Bayes BH procedure. Figure 1 displays the kernel smoothing density curves of the FDP and the NTD from several additional sets of simulations with large  $\varrho_2(\mathbf{P})$  and small  $\varrho_3(\mathbf{P})$  (transition matrices not shown). Consistent with the above results, the distribution of the NTD of the Bayes BH procedure is substantially more concentrated than that of the BH procedure.

### 3.4.2 Large $\varrho_2$ and relatively large $\varrho_3$

Table 3 displays the results of a set of simulations in which the parent Markov chain has a large  $\varrho_2(\mathbf{P})$ . However,  $\varrho_3(\mathbf{P})$  is also sampled to be relatively large. This makes the sampling of  $\mathbf{P}$  more difficult. In contrast, in the simulations of Section 3.4.1, there is actually no control on  $\varrho_3(\mathbf{P})$ . As

Table 2: Comparison of the Bayes BH and BH procedures again in the presence of large  $\varrho_2$  and small  $\varrho_3$ . The transition matrix of the parent Markov chain is (13) with  $\varrho_2 = 0.9566$ ,  $\varrho_3 = 0.1155$  and  $\psi = 10\%$ .

		BBH	BH	BBH	BH
$\epsilon$	$\alpha$	$\hat{\mu}_{\text{FDP}}$		$\text{SE}(\hat{\mu}_{\text{FDP}})$	
0.75	0.3	0.29201	0.3071	0.20283	0.20708
0.75	0.4	0.39533	0.39839	0.066775	0.13137
1	0.2	0.21889	0.20237	0.057124	0.096743
1	0.2	0.21572	0.19602	0.056001	0.094159
1	0.25	0.25369	0.2529	0.040186	0.062352
$\epsilon$	$\alpha$	$\hat{\mu}_{\text{NTD}}$		$\text{SE}(\hat{\mu}_{\text{NTD}})$	
0.75	0.3	5.365	10.233	3.1205	9.6933
0.75	0.4	36.139	35.77	7.9397	24.235
1	0.2	44.426	27.889	8.9892	17.99
1	0.2	42.673	26.994	8.8113	17.099
1	0.25	90.668	66.256	12.665	31.082

a result,  $\varrho_3(\mathbf{P})$  in those simulations is small and in most cases, the value is less than 0.3. The transition matrix of the parent Markov chain for Table 3 has  $\varrho_2(\mathbf{P}) = 0.9591$ ,  $\varrho_3(\mathbf{P}) = 0.5232$ ,

$$\mathbf{P} = \begin{pmatrix} 0.0014 & 0.9653 & 0.0124 & 0.0080 & 0.0129 \\ 0.9506 & 0.0011 & 0.0056 & 0.0155 & 0.0272 \\ 0.3236 & 0.1203 & 0.5312 & 0.0238 & 0.0011 \\ 0.8080 & 0.1589 & 0.0190 & 0.0025 & 0.0115 \\ 0.1898 & 0.7959 & 0.0053 & 0.0024 & 0.0066 \end{pmatrix} \quad (14)$$

with  $\boldsymbol{\pi} = (0.47319, 0.47681, 0.018907, 0.01173, 0.019363)^t$  and  $\psi = 0.05$ . Note that the proportion of false nulls is only half as large as the one in Section 3.4.1, making error control more difficult. Table 3 shows that in this case, even though the Bayes BH procedure has a harder time to control the FDR, it has substantially less variation in the FDP than the BH procedure does. Further, the power of the Bayes BH procedure is clearly superior to the BH procedure. For example, at a realized FDP  $\approx 0.35$ , the former makes about 204 true discoveries whereas the latter makes about 149. The Bayes BH procedure also has significantly lower SE-to-mean ratio of the NTD than the BH procedure. The phenomenon was repeatedly observed when  $\varrho_3(\mathbf{P})$  was sampled to be relatively large. Figure 2 provides further examples of the phenomenon. In contrast to Figure 1, the FDP density curves of the two procedures exhibit pronounced differences, and their NTD density curves are far apart instead of overlapping with each other.

As noted in the introduction, it is known that the BH procedure is often able to control the FDR under dependence, and certain properties of dependence are sufficient for this to happen [3, 10, 35, 36, 43]. Much less is known about the relationship between properties of dependence and other aspects of performance of multiple testing, such as stability and power. One advantage of random sampling from a large space of dependence structures is that it enables experimental investigation of the relationship. The properties of statistical dependence as characterized by  $\varrho_2(\mathbf{P})$  and  $\varrho_3(\mathbf{P})$  have been the focus here. The above results indicate that  $\varrho_2(\mathbf{P})$  alone is not sufficient to characterize the relative performances of the Bayes BH and the BH procedures. This numerical observation is consistent with what is now known about the mixing times of Markov

Table 3: Comparison of the Bayes BH and BH procedures in the presence of large  $\varrho_2$  and relatively large  $\varrho_3$ . The transition matrix of the parent Markov chain is (14) with  $\varrho_2 = 0.9591$ ,  $\varrho_3 = 0.5232$ , and  $\psi = 5\%$ .

		BBH	BH	BBH	BH
$\epsilon$	$\alpha$	$\hat{\mu}_{\text{FDP}}$		$\text{SE}(\hat{\mu}_{\text{FDP}})$	
1	0.2	0.29271	0.19	0.10223	0.23451
1	0.3	0.3676	0.30591	0.06527	0.19306
1	0.4	0.47105	0.40827	0.039127	0.11613
1.25	0.2	0.35611	0.1958	0.032559	0.086935
1.25	0.356	0.42145	0.35488	0.02103	0.041823
$\epsilon$	$\alpha$	$\hat{\mu}_{\text{NTD}}$		$\text{SE}(\hat{\mu}_{\text{NTD}})$	
1	0.2	23.739	9.0709	3.655	4.1191
1	0.3	53.006	13.288	11.094	9.1999
1	0.4	119.15	19.097	30.862	18.862
1.25	0.2	204.12	30.266	23.961	16.364
1.25	0.356	402.27	149.84	31.26	37.498

chains [11, 28]. The results in Table 3 and Figure 2 indicate that  $\varrho_3(\mathbf{P})$  may sometimes play a role in shaping the differences. In our simulations,  $\mathbf{P}$  is a  $5 \times 5$  matrix. Thus  $\varrho_2(\mathbf{P})$  and  $\varrho_3(\mathbf{P})$  together provide a significant part of information on the spectrum of the transition matrix. This points to a complex relationship between the dependence structure of the underlying hypotheses and the relative performances of the Bayes BH and the BH procedures.

### 3.4.3 Moderate $\varrho_2$

Table 4 displays results for the case of moderate value of  $\varrho_2(\mathbf{P})$ . In this set of simulations,  $\varrho_2(\mathbf{P}) = 0.6019$ ,

$$\mathbf{P} = \begin{pmatrix} 0.1999 & 0.6548 & 0.0537 & 0.0561 & 0.0354 \\ 0.8932 & 0.0765 & 0.0137 & 0.0088 & 0.0076 \\ 0.4431 & 0.4365 & 0.0721 & 0.0284 & 0.0199 \\ 0.1159 & 0.8271 & 0.0371 & 0.0129 & 0.0070 \\ 0.6080 & 0.0597 & 0.0505 & 0.0710 & 0.2108 \end{pmatrix} \quad (15)$$

with  $\boldsymbol{\pi} = (0.4969, 0.4031, 0.0376, 0.0349, 0.0275)^\top$  and  $\psi = 0.1$ . From the table, the NTD of the Bayes BH procedure is clearly more stable than that of the BH procedure. This can also be seen from the densities of the NTD shown in Figure 3. To a lesser extent, the FDP of the Bayes BH procedure is also more stable than that of the BH procedure except when  $\epsilon = 0.75$ ,  $\alpha = 0.3$  or  $0.4$ . In the latter situation, the variance of the FDP of the Bayes BH procedure is still competitive with that of the BH procedure.

### 3.4.4 Independence

Finally, as a test on the validity of the Bayes BH procedure, Table 5 displays results when  $\varrho_2(\mathbf{P}) = 0$  and, as a result, the entries of  $\boldsymbol{\eta}$  are i.i.d. In the simulations,  $\psi = 0.05$  and  $\mathbf{P} = \mathbf{1}_5 \boldsymbol{\pi}^\top$ , where the stationary probability vector  $\boldsymbol{\pi}$  is randomly sampled as in Section 3.1. As can be seen, under

Table 4: Comparison of the Bayes BH and BH procedures in the presence of moderate  $\varrho_2$ . The transition matrix of the parent Markov chain is (15) with  $\varrho_2 = 0.6019$  and  $\psi = 10\%$ .

		BBH	BH	BBH	BH
$\epsilon$	$\alpha$	$\hat{\mu}_{\text{FDP}}$		$\text{SE}(\hat{\mu}_{\text{FDP}})$	
0.75	0.3	0.29201	0.3071	0.20283	0.20708
0.75	0.4	0.22492	0.3014	0.2719	0.2124
0.75	0.5	0.4966	0.50118	0.03134	0.045887
1	0.2	0.192	0.2012	0.075112	0.084956
1	0.3	0.294	0.30074	0.035074	0.040486
$\epsilon$	$\alpha$	$\hat{\mu}_{\text{NTD}}$		$\text{SE}(\hat{\mu}_{\text{NTD}})$	
0.75	0.3	2.094	9.431	1.8608	9.2073
0.75	0.4	22.28	38.658	5.6593	25.412
0.75	0.5	121.26	138.26	12.793	60.366
1	0.2	22.542	32.796	5.8157	18.614
1	0.3	116.64	129.34	12.984	42.037

independence, both the Bayes BH procedure and the BH procedure are valid. In terms of power, although the NTD of the Bayes BH procedure has a smaller average than that of the BH procedure, its stability is clearly superior. Similar phenomenon can be seen from the density plot in Figure 4.

Table 5: Comparison of the Bayes BH and BH procedures when there is no dependence in data.

		BBH	BH	BBH	BH
$\epsilon$	$\alpha$	$\hat{\mu}_{\text{FDP}}$		$\text{SE}(\hat{\mu}_{\text{FDP}})$	
1	0.3	0.29028	0.29932	0.19437	0.18847
1	0.4	0.39102	0.39308	0.081393	0.123
1	0.5	0.50282	0.50157	0.039488	0.065857
1.25	0.2	0.19442	0.19951	0.072535	0.083292
$\epsilon$	$\alpha$	$\hat{\mu}_{\text{NTD}}$		$\text{SE}(\hat{\mu}_{\text{NTD}})$	
1	0.3	5.322	11.16	2.9969	9.6522
1	0.4	22.881	32.613	5.9599	19.899
1	0.5	79.951	87.779	10.701	36.661
1.25	0.2	24.838	31.121	5.9329	15.893

## References

- [1] BARVINOK, A. (2010). What does a random contingency table look like? *Combin. Probab. Comput.* **19**, 4, 517–539.
- [2] BENJAMINI, Y. AND HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300.

- [3] BENJAMINI, Y. AND YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 4, 1165–1188.
- [4] BORDENAVE, C., CAPUTO, P., AND CHAFAÏ, D. (2012). Circular law theorem for random Markov matrices. *Probab. Theory Related Fields* **152**, 3–4, 751–779.
- [5] BRUALDI, R. A., PARTER, S. V., AND SCHNEIDER, H. (1966). The diagonal equivalence of a nonnegative matrix to a stochastic matrix. *J. Math. Anal. Appl.* **16**, 31–50.
- [6] CAPPELLINI, V., SOMMERS, H.-J., BRUZDA, W., AND ŻYCKOWSKI, K. (2009). Random bistochastic matrices. *J. Phys. A* **42**, 36, 365209, 23 pp.
- [7] CHAFAÏ, D. (2010). The Dirichlet Markov ensemble. *J. Multivariate Anal.* **101**, 3, 555–567.
- [8] CHATTERJE, S., DIACONIS, P., AND SLY, A. (2010). Properties of uniform doubly stochastic matrices. *ArXiv e-prints*.
- [9] CHI, Z. (2011). Effects of statistical dependence on multiple testing under a hidden Markov model. *Ann. Statist.* **39**, 1, 439–473.
- [10] CLARKE, S. AND HALL, P. (2009). Robustness of multiple testing procedures against dependence. *Ann. Statist.* **37**, 1, 332–358.
- [11] DIACONIS, P. (1995). The cutoff phenomenon in finite Markov chains. *Proc Natl Acad Sci USA* **93**, 1659–1664.
- [12] DIACONIS, P., LEBEAU, G., AND MICHEL, L. (2012). Gibbs/Metropolis algorithms on a convex polytope. *Math. Z.* **272**, 1-2, 109–129.
- [13] EAVES, B. C., HOFFMAN, A. J., ROTHBLUM, U. G., AND SCHNEIDER, H. (1985). Line-sum-symmetric scalings of square nonnegative matrices. *Math. Programming Stud.* **25**, 124–141. Mathematical programming, II.
- [14] EFRON, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.* **102**, 477, 93–103.
- [15] FAN, J., HAN, X., AND GU, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *J. Amer. Statist. Assoc.* **107**, 499, 1019–1035.
- [16] FARCOMENI, A. (2008). A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Stat. Methods Med. Res.* **17**, 4, 347–388.
- [17] FRANKLIN, J. AND LORENZ, J. (1989). On the scaling of multidimensional matrices. *Linear Algebra Appl.* **114/115**, 717–735.
- [18] FRIGUET, C. (2012). A general approach to account for dependence in large-scale multiple testing. *Journal de la Société Française de Statistique* **153**, 2, 100–122.
- [19] FRIGUET, C., KLOAREG, M., AND CAUSEUR, D. (2009). A factor model approach to multiple testing under dependence. *J. Amer. Statist. Assoc.* **104**, 488, 1406–1415.
- [20] GENOVESE, C. AND WASSERMAN, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. Ser. B* **64**, 3, 499–517.

- [21] GOLDBERG, G. AND NEUMANN, M. (2003). Distribution of subdominant eigenvalues of matrices with random rows. *SIAM J. Matrix Anal. Appl.* **24**, 3, 747–761 (electronic).
- [22] GORDON, A., GLAZKO, G., QIU, X., AND YAKOVLEV, A. (2007). Control of the mean number of false discoveries, Bonferroni and stability of multiple testing. *Ann. Appl. Statist.* **1**, 1, 179–190.
- [23] HARTFIEL, D. J. (1971). Concerning diagonal similarity of irreducible matrices. *Proc. Amer. Math. Soc.* **30**, 419–425.
- [24] HARTFIEL, D. J. (1974). A study of convex sets of stochastic matrices induced by probability vectors. *Pacific J. Math.* **52**, 405–418.
- [25] KNIGHT, P. A. (2008). The Sinkhorn-Knopp algorithm: convergence and applications. *SIAM J. Matrix Anal. Appl.* **30**, 1, 261–275.
- [26] KÜNSCH, H., GEMAN, S., AND KEHAGIAS, A. (1995). Hidden Markov random fields. *Ann. Appl. Probab.* **5**, 3, 577–602.
- [27] LEEK, J. T. AND STOREY, J. D. (2008). A general framework for multiple testing dependence. *Proc Natl Acad Sci USA* **105**, 48, 18718–18723.
- [28] LEVIN, D. A., PERES, Y., AND WILMER, E. L. (2009). *Markov chains and mixing times*. American Mathematical Society, Providence, RI. With a chapter by James G. Propp and David B. Wilson.
- [29] MARSHALL, A. W. AND OLKIN, I. (1968). Scaling of matrices to achieve specified row and column sums. *Numer. Math.* **12**, 83–90.
- [30] OWEN, A. B. (2005). Variance of the number of false discoveries. *J. R. Stat. Soc. Ser. B* **67**, 3, 411–426.
- [31] QIU, X., KLEBANOV, L., AND YAKOVLEV, A. (2005). Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes. *Stat. Appl. Genet. Mol. Biol.* **4**, Art. 34, 32 pp. (electronic).
- [32] ROQUAIN, E. AND VILLERS, F. (2011). Exact calculations for false discovery proportion with application to least favorable configurations. *Ann. Statist.* **39**, 1, 584–612.
- [33] ROSENTHAL, J. S. (1995). Convergence rates for Markov chains. *SIAM Rev.* **37**, 3, 387–405.
- [34] ROTHBLUM, U. G. AND SCHNEIDER, H. (1989). Scalings of matrices which have prespecified row sums and column sums via optimization. *Linear Algebra Appl.* **114/115**, 737–764.
- [35] SARKAR, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist.* **30**, 1, 239–257.
- [36] SARKAR, S. K. (2006). False discovery and false non-discovery rates in single-step multiple testing procedures. *Ann. Statist.* **34**, 1, 394–415.
- [37] SARKAR, S. K., ZHOU, T., AND GHOSH, D. (2008). A general decision theoretic formulation of procedures controlling FDR and FNR from a Bayesian perspective. *Statistica Sinica* **18**, 925–945.



- [38] SIMES, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 3, 751–754.
- [39] SINKHORN, R. (1964). A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann. Math. Statist.* *35*, 876–879.
- [40] SINKHORN, R. (1967). Diagonal equivalence to matrices with prescribed row and column sums. *Amer. Math. Monthly* *74*, 402–405.
- [41] STOREY, J. D., TAYLOR, J. E., AND SIEGMUND, D. O. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc. Ser. B* **66**, 1, 187–205.
- [42] SUN, W. AND CAI, T. T. (2009). Large-scale multiple testing under dependence. *J. R. Stat. Soc. Ser. B* **71**, 2, 393–424.
- [43] WU, W.-B. (2008). On false discovery control under dependence. *Ann. Statist.* **36**, 1, 364–380.

Figure 1: Densities of the FDP (left) and NTD (right) of the Bayes BH and BH procedures in the presence of large  $\varrho_2$  and small  $\varrho_3$ .

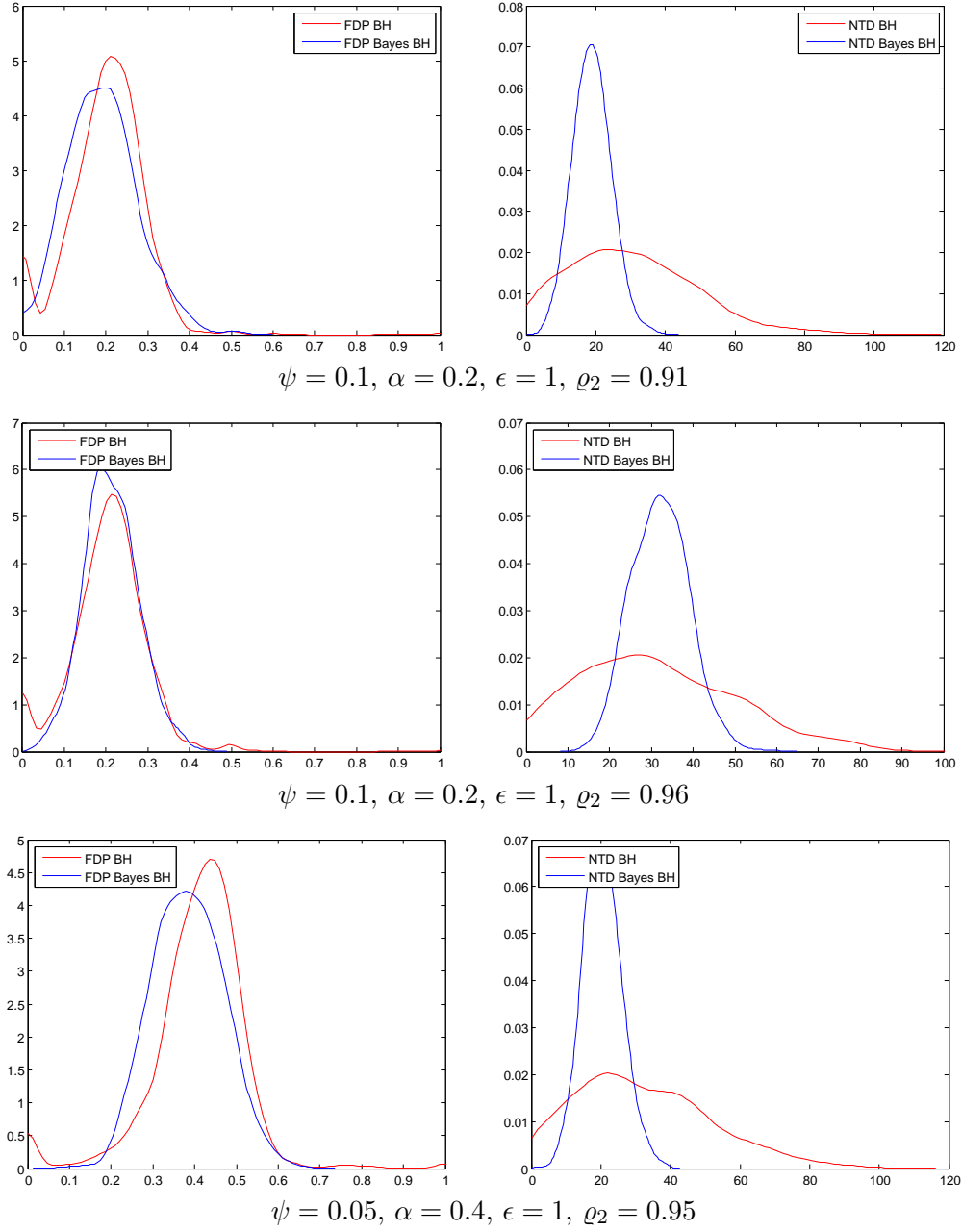


Figure 2: Densities of the FDP and NTD of the Bayes BH and BH procedures in the presence of large  $\varrho_2$  and relatively large  $\varrho_3$ .

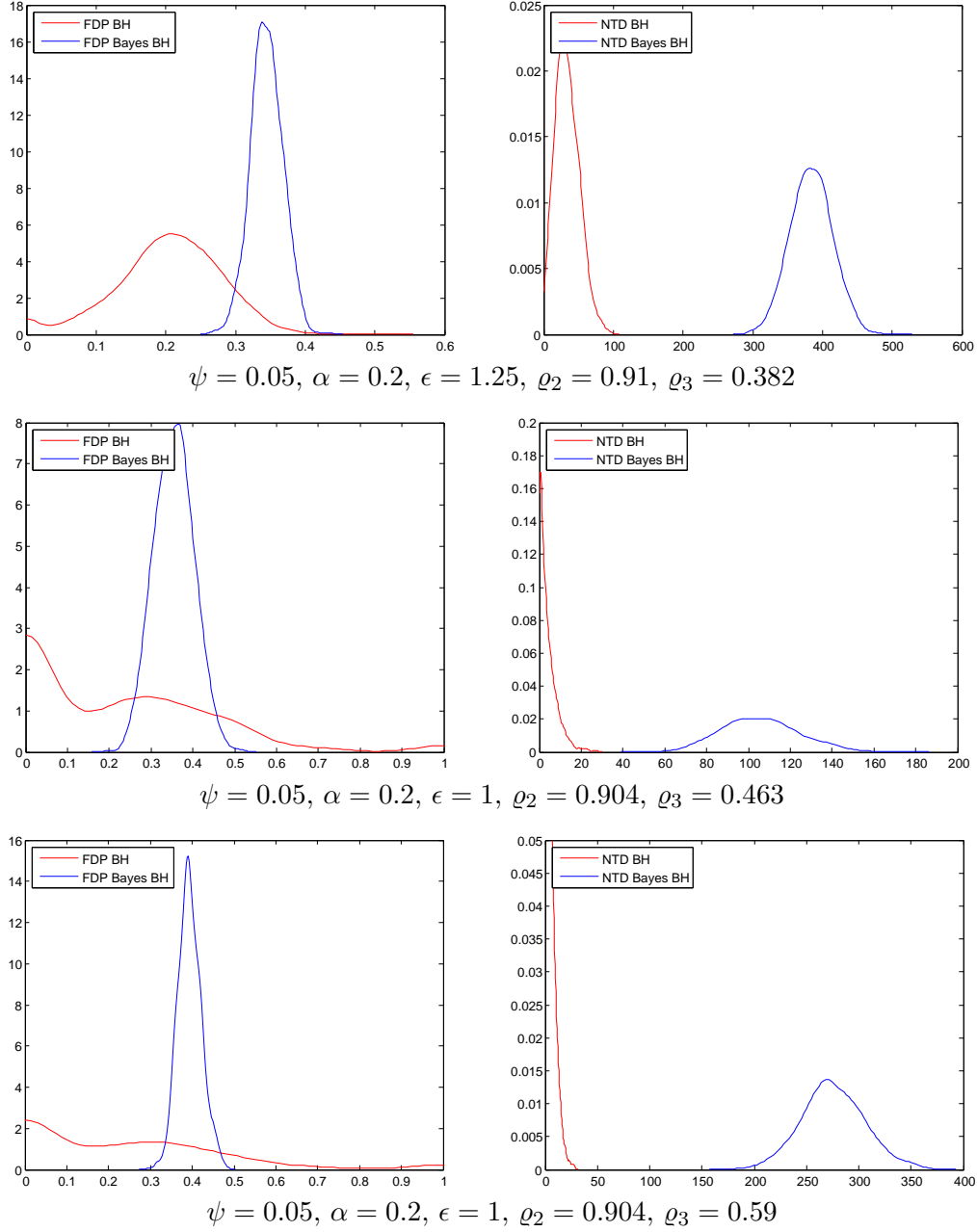


Figure 3: Densities of the FDP and NTD of the Bayes BH and BH procedures in the presence of moderate  $\varrho_2$ .

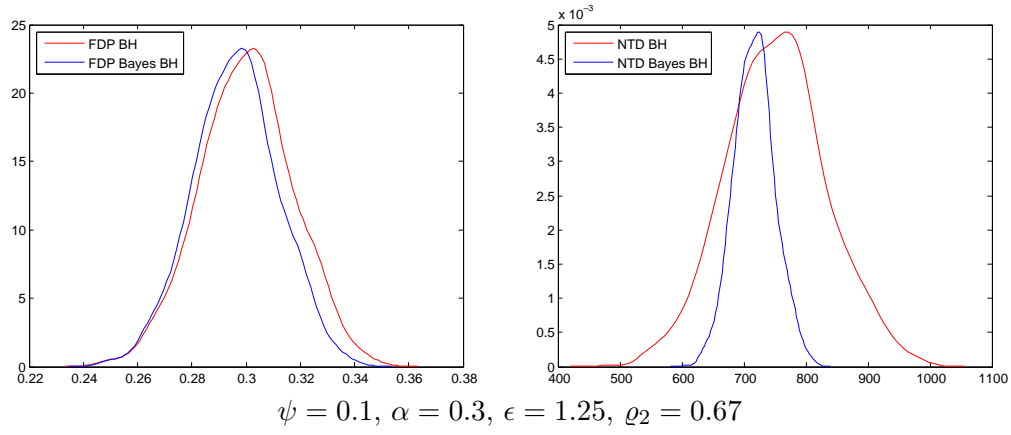


Figure 4: Densities of the FDP and NTD of the Bayes BH and BH procedures when hypotheses are independent.

